VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Topluluk Sınıflandırma Yöntemleri ve PCA Kullanarak Zararlı Url Tespiti

Kübra KÖKSAL [1*], Buket DOĞAN [2], Zehra Aysun ALTIKARDEŞ [2,3]

[1] *Marmara Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, 34722, Kadıköy/İstanbul, Türkiye*
[2] *Marmara Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, 34722, Kadıköy/İstanbul, Türkiye*
[3] *Marmara Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Türkiye, 34722, Kadıköy/İstanbul*

Makale Bilgisi

Başvuru:
06/07/2021
Kabul:
18/08/2021

Özet

Teknolojinin gelişmesi ve internet kullanıcı sayısındaki artışla orantılı olarak siber suçlarda da artış gözlemlenmiştir. Birçok farklı siber saldırı tekniği bulunmaktadır. Bu saldırı tekniklerinden biri olan kötü amaçlı web siteleri, siber saldırılar ve dolandırıcılık olaylarında önemli rol oynamaktadır. İnternette masum görünen bir bağlantıya tıklamak veya e-posta ve mesaj yoluyla gönderilen bir web sayfasını ziyaret etmek arka planda kimlik avı kampanyalarının başlatılmasına, kötü amaçlı yazılımların, casus yazılımların, fidye yazılımların indirilmesine ve ciddi parasal kayıplar oluşmasına yol açar. Dolayısıyla bu tehditlerin etkin bir şekilde tespit edilmesi ve önlenmesi bireyler, kurumlar ve hükümetler için oldukça önemli bir konu haline gelmiştir. Kara listeye dayalı yöntemler, kötü amaçlı URL'leri tanımlamak için kullanılan standart yöntemlerden biridir. Ancak kara listeler hiçbir zaman kapsamlı değildir ve yeni oluşturulan URL'leri algılama yeteneğinden yoksundur. Kara listeye dayalı yöntemlerin mevcut ihtiyacı ve eksiklikleri de göz önünde bulundurularak bu çalışmada toplulukla öğrenme yöntemleri kullanılarak bir sınıflandırma yaklaşımı önerilmiştir. Çalışmada iyi huylu ve kötü huylu URL'lerden elde edilmiş 79 sözcüksel özellik içeren Kanada Siber Güvenlik Enstitüsü'nün URL veriseti (ISCX-URL-2016) üzerinde çalışılmıştır. Verisetinde benign, spam, phishing, malware ve defacement olmak üzere beş farklı URL türü bulunmaktadır. Toplam 7781 iyi huylu ve 28.917 tane zararlı URL kaydı üzerinde zararlı, zararsız etiketleri kullanılarak ikili sınıflandırma işlemi ve beş farklı etiket bilgisi kullanılarak çoklu sınıflandırma işlemi gerçekleştirilmiştir. Makine öğrenmesi yöntemlerinden Rastgele Orman algoritması uygulanan yöntemin başarısının sınanması için 10-katlamalı çapraz doğrulama (10-fold cross validation) ile birlikte kullanılmıştır ve 10 temel bileşen kullanılarak ikili sınıflandırma problemi için ortalama %99.42, çoklu sınıflandırma problemi için ortalama %95.68 doğruluk değeri elde edilmiştir. Böylece her gün yeni web sitelerinin katıldığı bu dinamik internet ağını kötü niyetli tasarlanmış web sitelerinden korumaya yönelik yüksek başarım oranına sahip bir model önerisi sunulmuştur.

Anahtar Kelimeler: *Kötü niyetli URL, siber güvenlik, makine öğrenmesi, sıradışı veri, rastgele orman*

[1]* İletişim e-posta: kubra.koksal@marun.edu.tr

Malicious url detection using ensemble learning methods and PCA

Abstract

In parallel with the development of technology and the increase in the number of internet users, an increase in cybercrime has been observed. There are many different cyberattack techniques. Malicious websites, one of these attack techniques, play an important role in cyberattacks and fraud events. Clicking on an innocent-looking link on the Internet or visiting a web page sent via email or text will result in phishing campaigns being launched in the background, downloading malware, spyware, ransomware, and serious monetary losses. Therefore, effective detection and prevention of these threats has become a very important issue for individuals, institutions and governments. Blacklist-based methods are one of the standard methods used to identify malicious URLs. However, blacklists are never comprehensive and lack the ability to detect newly created URLs. Considering the current needs and deficiencies of blacklist-based methods, a machine learning based classification approach was used in this study to combat malicious URLs. In the study, the URL data set of the Canadian Cyber Security Institute (ISCX-URL-2016) was studied, which contains 79 lexical features obtained from benign and malignant URLs. There are five different URL types in the dataset: benign, spam, phishing, malware and defacement. A binary classification process using harmless, malicious labels and a multi-classification process using five different labels information was performed on a total of 7781 benign, harmless and 28,917 malicious URL records. Random Forest algorithm, one of the machine learning methods, used together with 10-fold cross validation to validate the success of the applied method, and an average accuracy value of 99.42% for the binary classification problem and 95.68% for the multiple classification problem was obtained. Thus, a model proposal with a high-performance rate is presented to protect this dynamic internet network, where new websites are added every day, from maliciously designed websites.

Keywords: *Malicious URL, cyber security, machine learning, outlier data, random forest*

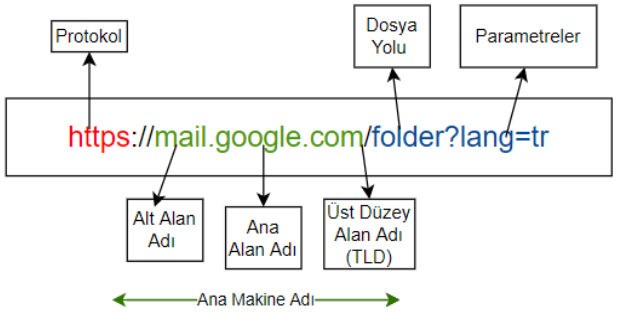
1 Giriş

Teknolojideki ilerlemeler toplumu daha fazla çevrimiçi yapılara yönlendirmiştir. Çevrimiçi ortamda daha fazla kişi ve işletme olduğunda, internette ve bilgisayarlarımızda saklanan özel bilgileri korumak zordur. Teknoloji sürekli gelişmekte ve değişmektedir, bu nedenle güvenlik önlemleri kullanıcıları korumaya devam etmek için uyarlanmalıdır.

Siber güvenlik, güvenli sistemler oluşturmak ve mevcut sistemleri siber suçlara karşı güvence altına almak amacıyla oluşan bilgisayar bilimi içerisindeki bir alandır. Siber saldırılar "bir varlığı ifşa etme, değiştirme, devre dışı bırakma, imha etme, çalma, yetkisiz erişim elde etme veya yetkisiz bir şekilde kullanma girişimi" olarak tanımlanır ve kişiye büyük miktarlarda zarar veren, en hızlı büyüyen, gelişen suçlardan biridir. Bir varlık, bir kuruluş veya birey için değerli bir şeydir. Yeni yazılım geliştirildikçe, yararlanabilecek yeni güvenlik açıkları potansiyeli vardır. Bu sorunla mücadele etmek ve saldırıları önlemek, sistemlerin ve ağların daha iyi izlenmesini gerektirir. Bu, siber suç faaliyetlerini bulmak için halka açık verileri izleyerek potansiyel saldırıları önleyen araştırmalar ortaya çıkmıştır. Sisteme zarar vererek kişisel bilgilere erişilebilmesini sağlayan mesaj, e-posta yoluyla gönderilen zararlı bağlantılar izleme gerektiren halka açık veri kaynaklarına örnek olarak verilebilir [1]. Siber saldırılar maliyet, altyapı ve verilere verilen zararı, hırsızlığı, sahtekârlığı ve üretkenliği kaybetmeyi içerir, ancak bunlarla sınırlı değildir.

İnternet Kullanımı ve Dünya Nüfus İstatistikleri (Internet Usage and World Population Statistics) tahminlerine göre 2021 yılında yedi milyardan fazla İnternet kullanıcısı bulunmaktadır ve bu da siber suç faaliyetlerini izlemeyi gittikçe zorlaştırır. Bu durum birçok şirket için büyük bir endişe ve öncelik haline gelmiştir. Birçok şirket ve birey kendini siber saldırılara karşı korumak için siber güvenlik alanında giderek yüksek miktarda harcamalar yapmıştır. Gartner, Inc.'in son tahminine göre risk yönetimi ve güvenlik alanındaki harcamalar 2020 yılında %6.4 arttı ve dünya çapında bu teknolojilere ve hizmetlere yapılan harcamaların 2021'de 150,4 milyar dolara ulaşarak %12,4 artması bekleniyor [2].

Milyarlarca internet kullanıcısı arasında, siber suçlular olarak bilinen başkalarından yararlanan insanlar bulunmaktadır. Bu saldırganlar kimliklerini ve konumlarını gizlemek için büyük çaba sarf eder. Çevrimiçi takma adlar ve farklı konumlar kullanılması bu saldırganların izlenmesi zorlaştırır ve saldırıları karmaşık olabilir. Saldırı yollarından biri, Tekdüzen Kaynak Konum Belirleyicilerinin (Uniform Resource Locator-URL) kötüye kullanılmasıdır [1]. Bu çalışmada da zararlı URL yapıları kullanılarak yapılan siber saldırıların toplulukla öğrenme yöntemleri kullanılarak tespit edilmesi hedeflenmiştir. Şekil 1 örnek bir URL yapısını göstermektedir. Bir URL protokol, alan adı ve dosya yolu olmak üzere temelde üç bölümden oluşmaktadır. Çalışmada iyi huylu ve kötü huylu URL'lerden elde edilmiş 79 sözcüksel özellik içeren bir veriseti kullanılmıştır. PCA boyut azaltma tekniği kullanılarak model oluşturulması için gereken zaman ve depolamadan avantaj sağlanırken eşdoğrusallığın giderilmesi ile makine öğrenmesi modelinin performansının arttırılması hedeflenmiştir.



Şekil 1. Örnek Bir URL Yapısı

Toplulukla öğrenme yöntemlerinden Rastgele Orman, AdaBoost ve Gradyan Arttırma kullanılarak üç farklı model oluşturulmuştur ve sonuçlar doğruluk, model oluşturma süresi gibi parametreler açısından değerlendirilmiştir. Bu çalışmanın ikinci bölümünde zararlı url tespiti ile ilgili literatürdeki önemli çalışmalar incelenmektedir, üçüncü bölümde önerilen yöntemin işlem adımları sunulmaktadır ve son bölümde elde edilen bulgular ve sonuçlar tartışılmaktadır.

2 İlgili çalışmalar

Bu kısımda zararlı URL tespiti ile ilgili yapılmış literatürdeki önemli çalışmalara yer verilmektedir.

McGrath vd. URL ve alan uzunluğu gibi özellikleri kullanarak iyi huylu ve kimlik avı URL'leri arasındaki farkları analiz etmiştir. Kimlik avı URL'lerinin anatomisini, etki alanlarının kaydını ve bu siteleri barındırmak amacıyla kullanılan makineleri inceleyerek kimlik avı amacıyla kullanılan alan adlarının farklı uzunluklara ve farklı konumlara sahip olduğunu göstermiştir [3]. Ma vd. kimlik avı web sitelerindeki etki alanlarının davranışlarını temel alarak sözlü ve kara listeye alınmış ana bilgisayar tabanlı özellikler üzerinden şüpheli URL'leri %99 doğruluk değeri ile tanımlayan bir model geliştirmiştir [4].

Thomas, K. ve ark., twitter ve e-postalardaki spam iletilerinden toplanan kötü amaçlı URL'leri tanımlamak için sözcüksel, ana bilgisayar ve sayfa içeriğine dayalı özellikler kullanmıştır [5]. Choi vd. kötü amaçlı URL'leri tespit etmek ve spam, kimlik avı ve kötü amaçlı yazılım gibi saldırı türlerini belirlemek için bir makine öğrenme yöntemi sunmuştur. Çalışmaları, sözlük, bağlantı popülerliği, web sayfası içeriği, DNS, DNS sabitleme ve ağ izleme gibi altı farklı alandaki özellikleri kullanan birden fazla kötü amaçlı URL türünü içerir. Ancak, elde edilen sonuçlar, sözcük özelliklerini kullanmanın spam ve malware URL veri kümesi için daha düşük doğruluk sağladığını göstermektedir [6].

Lin, M. vd. sözcük özelliklerini tamamlamak için URL'nin açıklayıcı özelliklerini kullanmışlardır. Kötü amaçlı URL'leri sınıflandırmak için URL dizesinin sözcük bilgilerini ve statik özelliklerini birleştirmişlerdir. Ana makine ve içerik tabanlı analiz olmadan, bu deneyde, iki dakika içinde iki milyon URL ile başa çıkabilmişler ve önerilen yöntemleri kötü amaçlı örneklerin yaklaşık %9'unu kaçırmıştır [7].

Bilinen korumalı web siteleri ile makine öğrenimi tabanlı kimlik avı tespiti Chu vd. tarafından gerçekleştirilmiştir [8]. Yalnızca sözcüksel ve alan adı özelliklerine dayanarak yapılan çalışmada, %91'in üzerinde tespit oranına ulaşılmıştır.

Mohammad Saiful Islam Mamun vd. verisetinin ayrıntılarını ve özelliklerini özetleyen bir araştırma raporu yayınlamıştır. Bu çalışma kapsamında CfsSub Infogain özellik seçme algoritmaları Rastgele Orman yöntemi ile birlikte kullanılarak ikili sınıflandırma için %99 ve çoklu sınıflandırma için %93-99 doğruluk oranı elde edilmiştir [9].

Apoorva Joshi vd. sözcüksel özellikler ile makine öğrenmesi tekniklerini kullanarak zararlı URL'lerin sınıflandırılması üzerine çalışmışlardır. Farklı algoritmalar arasından %99 değeri ile en yüksek doğruluk değerini Rastgele Orman sınıflandırma yöntemi kullanarak elde etmişlerdir [10].

Alexander Powell vd. bu çalışmada da kullanılan veriseti üzerinde çeşitli özellik çıkarma ve makine öğrenmesi tekniklerini kullanarak özellik seçme işleminin algoritma performansı ve sınıflandırma süresi üzerindeki etkisini incelemiştir. Üç farklı veriseti üzerinde yapılan analizlerde ISCX-URL-2016 veriseti üzerinde Karar Ağacı ve Ekstra Ağaç Sınıflandırıcı (ExtraTree Classifier) kullanılarak oluşturulan model %99 değeri en yüksek doğruluk oranına ulaşmıştır [11].

Siddharth Singhal vd. kötü niyetli ve iyi niyetli URL'Lerden sözcüksel, ana bilgisayar tabanlı ve içerik tabanlı özellikleri toplamıştır. Rastgele ormanlar, Gradyan Artırılmış Karar Ağaçları ve Derin Sinir Ağı sınıflandırıcıları kullanılarak gerçekleştirilen sınıflandırma sonuçları Gradyan Artırılmış Karar Ağaçları algoritmasının %96.4 değeri ile yüksek bir doğruluğa ulaştığını göstermektedir [12]. Shuai Wang vd. ISCX-URL-2016 veriseti üzerinde Naive Bayes algoritmasını kullanarak bir ikili sınıflandırma modeli oluşturmuştur. Özellik seçme algoritması sonucunda seçilen üç özellik kullanılarak oluşturulan model %90.18 doğruluk değerine ulaşmıştır [13].

Ozgur Koray Sahingoz vd. yedi farklı sınıflandırma algoritması ile birlikte doğal dil işleme (NLP) tabanlı özellikleri kullanarak gerçek zamanlı anti-phishing sistemini önermişlerdir. Uygulanan sınıflandırma algoritmalarından elde edinilen sonuçlara göre NLP tabanlı Rastgele Orman algoritması %97.98 doğruluk oranı ile en iyi performansı göstermiştir [14].

Literatür araştırması sonucunda sözcüksel özellikleri ve özellik seçme algoritmalarını kullanılarak yüksek doğruluk değerlerine ulaşan modellerin olduğu görülmüştür.

Bu çalışmada da bir boyut azaltma yöntemi olan PCA'nın sözcüksel özellikler üzerinde kullanılması ile sınıflandırma işleminin yapılması hedeflenmiştir.

3 Materyal ve yöntem

Çalışmada zararsız ve zararlı URL adreslerinin bulunduğu, Kanada Siber Güvenlik Enstitüsü'nün URL veriseti (ISCX-URL-2016) kullanılmıştır [15]. Verisetinde benign, spam, phishing, malware ve defacement olmak üzere temel olarak beş farklı URL türü bulunmaktadır. Çalışmada bu URL'ler kullanılarak elde edilen sözcüksel özelliklerden oluşan veriseti kullanılmıştır. Hedef sütunla beraber toplam 80 sözcüksel özellik bulunmaktadır. Tablo 1'de görüldüğü gibi 7781 iyi huylu ve 28.917 tane zararlı URL kaydı ile birlikte toplamda 36698 kayıt vardır. Zararlı URL'ler de kendi içinde dört türe ayrılmaktadır.

Tablo 1. Url türlerine göre kayıt sayıları

URL türü	Kayıt sayısı
benign	7781
defacement	7931
phishing	7577
spam	6698
malware	6711

Sözcüksel özellikler URL'nin ana bilgisayar adı uzunluğu, URL uzunluğu, URL'de bulunan simgeler gibi metinsel özellikleridir. Verisetindeki tüm özelliklerin açıklamaları aşağıdaki gibidir [8] :

Entropi alan adı ve uzantısı: Kötü amaçlı web siteleri, meşru görünmesi için genellikle URL'ye ek karakterler ekler. Örneğin, CITI ifadesindeki son harf olan I değeri, 1 sayısı ile değiştirilerek CIT1 olarak yazılabilir. Karakter ekleme sonucunda URL'in, alan adının, dosya yolunun entropisi normalden daha fazla değişir. Rastgele oluşturulan kötü amaçlı URL'leri tanımlamak için alfabe entropisi kullanılır.

Character continuity rate: Karakter Süreklilik Oranı, abc567ti = (3 + 3 + 1) / 9 = 0 .77 örneğinde olduğu gibi, etki alanındaki her karakter türünün en uzun belirteç uzunluğunun toplamını bulmak için kullanılır. Kötü amaçlı web siteleri, değişken sayıda karakter türüne sahip URL'ler kullanır. Karakter süreklilik oranı sütunu harf, rakam ve sembol karakterlerinin sırasını belirler. Bir karakter türünün en uzun token uzunluğunun toplamı URL'nin uzunluğuna bölünür.

Uzunluk oranı ile ilgili özellikler: URL bölümlerinin (bağımsız değişken, yol, alan adı, URL) uzunluk oranı, anormal kısımları bulmak için hesaplanır. Örneğin argPathRatio (bağımsız değişken ve yol oranı), argUrlRatio (bağımsız değişken ve URL oranı), argDomainRatio (bağımsız değişkenin etki alanına bölünmesi), domainUrlRatio (etki alanının URL'ye bölünmesi), pathUrlRatio (Yolun URL'e bölünmesi), PathDomainRatio (Yolun etki alanına bölünmesi).

Tld (top level domain): Bazı kimlik avı(phishing) URL'leri, bir alan adı içinde birden çok üst düzey alan adı (https://www.example.gov.tr) kullanır.

Number Rate of DirectoryName, FileName, Domain, AfterPath, URL: Dizin adı, etki alanı, dosya adı, URL'in kendisi ve yoldan sonrası gibi URL

bölümlerindeki sayı oranını gösterir.*Uzunluk ile ilgili özellikler*: Değişkenlerin eklenmesi nedeniyle URL'nin uzunluğu uzar. URL Uzunluğu (url Len), alan adı uzunluğu (domain Len) ve dosya adı uzunluğu (File Name Len), bağımsız değişkenlerin en uzun kelime uzunluğu, en uzun yol belirteç uzunluğu, ortalama yol belirteci uzunluğu (avgpathtokenlen) gibi değerler örnek olarak verilebilir.

Ldl getArg: Kimlik avı URL'lerinde maskeleme, harflere rakam ekleyerek yapılır. Bu aldatıcı URL'lerin tespiti için, URL ve yoldaki harf rakam harf dizisi hesaplanır.

Harf, token ve sembol sayımı ile ilgili özellikler: URL'deki karakterlerin sıklığı harf, belirteç ve sembol şeklinde hesaplanır. Bu karakterler, URL'lerin şu bileşenlerinden kategorize edilir ve sayılır:

- *Symbol Count Domain:* Etki alanından: //.: /? =,; ()] + gibi semboller hesaplanır. Kimlik avı (Phishing) URL'leri iyi huylu olanlara göre daha fazla nokta içermektedir.
- *Domain token count:* Jetonlar URL Dizesinden alınır. Kötü Amaçlı URL'ler birden çok alan adı jetonu kullanır. Alanlardaki jeton sayısı hesaplanır.
- *Query Digit Count:* URL'nin sorgu bölümündeki sayı değerlerinin sayısı.

SpcharUrl: URL'ler, // gibi şüpheli olan özel karakterler kullanır.

Önerilen çalışma önişleme, boyut azaltma, toplulukla öğrenme yöntemleri kullanılarak sınıflandırma modellerinin oluşturulması ve sonuçların değerlendirilmesi aşamalarından oluşmaktadır. Yapılan çalışmaya ait akış diyagramı Şekil 2'de gösterilmiştir.

3.1 Ön işleme

Sınıflandırma performansını arttırmak için ilk olarak veri üzerinde ön işleme adımları uygulanmıştır.

3.2 PCA Boyut Azaltma

Temel bileşen analizinin (PCA) [16] ana fikri, veri kümesinde mevcut olan varyasyonu korurken, birbirleriyle ilişkili çok sayıda değişkenden oluşan bir veri kümesinin boyutluluğunu maksimum düzeyde indirmektir. Değişkenleri temel bileşenler olarak bilinen ve ortogonal olan yeni bir değişken grubuna dönüştürülerek yapılır. PCA ile boyut azaltma işleminde beş temel kavram bulunmaktadır.

Korelasyon: İki değişkenin birbiriyle ne kadar güçlü ilişkili olduğunu gösterir. -1 ile +1 arasında değer alır.

Ortogonal: Birbiriyle ilintisiz, yani herhangi bir değişken çifti arasındaki korelasyonun 0 olduğu durumdur.

Öz Vektör (Eigen Vector): Bir vektör dönüşüme uğradıktan sonra vektörün boyutundaki değişimden bağımsız olarak yönü aynı şekilde kalıyorsa bu dönüşüm vektörü öz vektör olarak isimlendirilir.

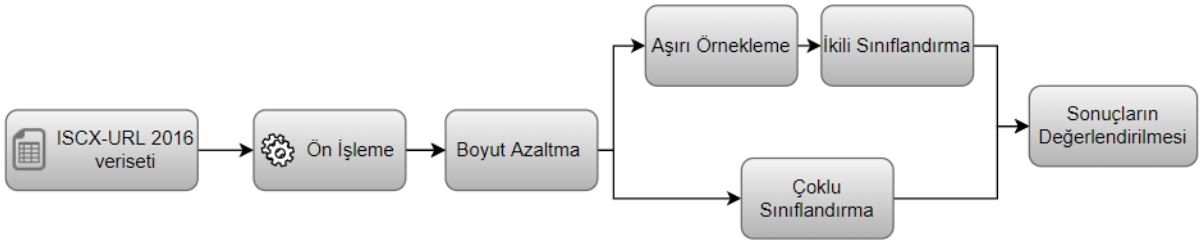
Öz Değer (Eigen Value): Yönü sabit olarak kalan fakat boyutu değişen bu öz vektörün uğradığı değişim işlemi sayısal olarak hesaplanabilir ve bu hesaplanan sayısal uzunluk değeri öz değer olarak isimlendirilir.

Kovaryans Matrisi: Kovaryans matrisi, değişken çiftleri arasındaki kovaryanslardan oluşur. (i, j) elemanı i ve j değişkeni arasındaki kovaryanstır.

Temel Bileşen Analizinin Adımları

1. Verilerin normalleştirilmesi. Bu, ortalaması sıfır olan bir veri kümesi üretilmesini sağlar.
2. Kovaryans matrisinin hesaplanması.
3. Özdeğer ve özvektörlerin hesaplanması, λ bu karakteristik denklemin çözümü ise bu değer A için bir özdeğerdir.

$$\det(\lambda I - A) = 0 \quad (1)$$



Şekil 2. Çalışmanın akış diyagramı

ƛ, A ile aynı boyuttaki bir birim matrisidir ve "det" matrisin determinantını belirtmektedir. Her bir özdeğer ƛ için, karşılık gelen bir öz vektör v değeri formül 2 kullanılarak bulunabilir.

$$(ƛI - A)\, v = 0 \quad (2)$$

4. Bileşenleri seçme ve bir özellik vektörü oluşturmada, özdeğerler en büyükten en küçüğe doğru sıralanır. N değişkenli bir verisetimiz varsa, karşılık gelen N özdeğere ve özvektöre sahip oluruz. En yüksek öz değere karşılık gelen özvektör, veri kümesinin ana bileşeni (principal component) dir. Belirlenen sayı doğultusunda özvektör seçilerek geri kalanları göz ardı edilir. Ardından, seçilen özvektörler kullanılarak özellik vektörü oluşturulur.
5. Ana Bileşenlerin Oluşturulması: Özellik vektörü tutmayı seçtiğimiz öz vektörleri kullanarak oluşturulan matris ve ölçeklendirilmiş veri, orijinal veri kümesinin ölçeklendirilmiş halidir.

$$(\text{ÖzellikV.})^{T} * \text{ÖlçeklendirilmişVeri}^{\text{Transpoz}} \quad (3)$$

Çalışmada kullanılan veri setinde 79 özellik bulunmaktadır ve bu özellikler üzerinde boyut azaltma yöntemlerinden Temel Bileşenler Analizi (PCA) kullanılarak daha iyi bir sınıflandırma modeli oluşturulması amaçlanmıştır. Boyut azaltma işlemi WEKA uygulaması kullanılarak gerçekleştirilmiştir. Weka uygulaması veriseti üzerinde boyut azaltma, özellik seçme ve sınıflandırma işlemlerinin kolayca gerçekleştirilebilmesini sağlayan bir veri madenciliği programıdır. Weka üzerinde PCA analizi için varsayılan ayar bilgileri kullanıldığında 21 temel bileşen oluşmaktadır ve bileşenler ile açıklanan varyans değeri %95'dir. Oluşan temel bileşenlerin sırayla açıklanan varyans değerine etkisi incelendiğinde 10 temel bileşenden sonraki bileşenlerin varyans değerine katkısının

$$\begin{aligned}&-0.182\text{ArgLen}-0.181\text{argDomanRatio}-\\&0.181\text{Entropy\_Afterpath}-\\&0.181\text{NumberRate\_DirectoryName}-\\&0.181\text{NumberRate\_Extension}-\\&0.181\text{NumberRate\_FileName}-\\&0.181\text{Querylength}-\\&0.181\text{avgpathtokenlen}-\\&0.181\text{Entropy\_Filename}-\\&0.181\text{NumberRate\_AfterPath}\end{aligned} \quad (4)$$

0.01626'dan daha düşük değerler olduğu ve açıklanan varyans değerine katkılarının çok az sonucuna varılmıştır.

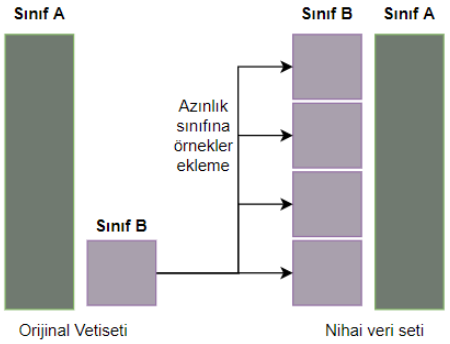
Bundan dolayı sadece ilk 10 temel bileşenin analiz kapsamında alınmasına ve elde edilen sonuçların 21 temel bileşen ile elde edinilen doğruluk değerleri ile karşılaştırılmasına karar verilmiştir. 10 temel bileşen kullanılarak açıklanan varyans değeri yaklaşık %80'dir.

Formül 4 açıklanan varyans değerine en yüksek katkıyı sağlayan ilk temel bileşeni göstermektedir.

3.3 Toplulukla Öğrenme Yöntemleri

İkili sınıflandırma işlemi için tüm iyi huylu zararsız kayıtlara hedef değer olarak 0 atanırken, zararlı sınıfına giren dört farklı URL sınıfına da 1 değeri atanmıştır.

Sonuçta oluşan verisetinde zararlı URL kayıt sayısı zararsız URL kayıt sayısına göre çok fazla olduğundan dolayı aşırı örnekleme (oversampling) işlemi uygulanmıştır. Bu işlem sonucunda iyi huylu URL kayıt sayısı zararlı URL kayıt sayısı ile aynı sayıya çıkartılmıştır. Şekil 3'te aşırı örnekleme işlemi gösterilmiştir. Çoklu sınıflandırmada toplamda beş sınıf olduğundan dolayı her bir sınıfa rastgele olarak 0-4 arasındaki değerler atanarak hedef değerin sayısallaştırılması işlemi gerçekleştirilmiştir.



Şekil *3.* Aşırı örnekleme işlemi

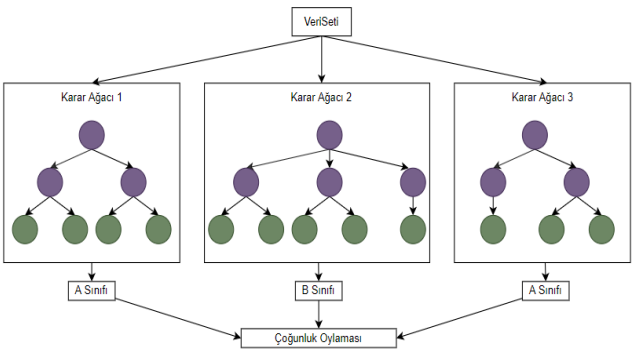
Çalışma kapsamında ikili ve çoklu sınıflandırma modellerinin oluşturulması için toplulukla öğrenme yöntemleri seçilmiştir.

İlgili çalışmalarda zararlı URL tespitinde sıklıkla kullanılan ve yüksek doğruluk değerlerine ulaşan Rastgele Orman algoritması PCA boyut azaltma yöntemi ile birlikte kullanılarak oluşturulan model AdaBoost ve Gradient Boosting gibi farklı toplulukla öğrenme yöntemleri ile ve aynı veriseti üzerinde farklı yöntemler uygulayan çalışmalar ile karşılaştırılmıştır.

Toplulukla öğrenme, sınıflandırma performansını arttırmak için birden fazla öğrenici kullanılarak model oluşturulan bir makine öğrenmesi yöntemidir. Torbalama ve yükseltme olmak üzere iki farklı şekilde uygulanabilir.

3.3.1 Torbalama (Bagging)

Torbalama rastgele verisetleri oluşturularak elde edinilen modelleri paralel olarak eğiten bir makine öğrenmesi yöntemidir. Rastgele Orman torbalama yöntemlerine örnek olarak verilebilir ve bu çalışmada da Rastgele Orman yöntemi kullanılarak bir zararlı URL tespit modeli oluşturulmuştur. Rastgele orman algoritması [17], karar ağacına dayanan bir toplulukla öğrenme yöntemidir. Karar ağaçlarının veriyi aşırı öğrenme probleminin üstesinden gelinmesini sağlayan bir yapıya sahiptir. Sınıflandırma işlemi için birden fazla karar ağacı yapısını kullanarak sınıflandırma değerini yükseltir. Örneğin N eğitim seti olduğunu ve her ağacın rastgele bir alt eğitim seti olarak N boyutunda eğitim örneği seçtiğini varsayalım.



Şekil 4. Rastgele Orman Algoritması

Verisetinde M özellik varsa, m (m <M) boyutunda olacak şekilde her bir karar ağacı bir alt özellik vektörü oluşturur ve ardından bu alt özellik kümesi kullanılarak karar ağacı budama işlemi yapılmadan oluşturulur. Bu şekilde, her ağaç farklı alt eğitim setlerine göre eğitim sonuçları elde etmiş olur. Rastgele Orman modeline verilen bir örnek için her bir karar ağacı tarafından ayrı ayrı değerlendirilir ve nihai sınıflandırma oylama sonuçlarına göre belirlenir; yani, birkaç zayıf sınıflandırıcının sonuçları güçlü bir sınıflandırıcı oluşturmak için birleştirilir. Son aşamada karar ağaçlarından elde edilen oylama sonuçları çoğunluk oylaması yöntemi kullanılarak değerlendirilir ve modele sunulan örnek üzerinde nihai karara varılmış olur. Şekil 4'te Rastgele Orman algoritmasının çalışma adımları gösterilmektedir.

3.3.2 Yükseltme (Boosting)

Yükseltme yöntemleri torbalama yöntemlerinin aksine rastgele olarak oluşturulan verisetleri üzerinde sıralı bir eğitim işlemi gerçekleştirmektedir. Paralel öğrenmede her bir model diğer modellerin sonucundan bağımsız olarak eğitim işlemini gerçekleştirirken sıralı öğrenmede her model bir önceki model sonucunu da dikkate almaktadır. Doğru ve yanlış olarak sınıflandırılan örneklerin ağırlıkları elde edinilen sonuçlara göre değiştirilir. AdaBoost, Gradyan Artırılmış Karar Ağaçları çalışma kapsamında incelenen yükseltme toplulukla öğrenme yöntemleridir. AdaBoost modeldeki her bir karar ağacını belirleme açısından Rastgele Orman algoritmasına benzemesine rağmen ağaç büyüklüklerinde bir kısıtlama mevcuttur. AdaBoost bir node ve iki yapraklı ağaçlardan oluşur. Ayrıca Torbalama yöntemlerinde hangi ağacın önce oluştuğu önemli değilken AdaBoost'ta bir sonraki düğümlerin oluşmasını etkiler. Gradyan Artırılmış Karar Ağaçları da Rastgele Orman algoritmasına benzer modeller oluşturan bir yöntemdir.
Ancak AdaBoost gibi her ağaçtan sonra iyileştirme yapmak için bir düğüm oluşturmak yerine yaprak düğüm ile başlar ve ardından Gradyan Artırılmış karar ağacını oluşturur. Oluşturulan tüm modellerin doğruluk değerleri 10-katlamalı çapraz doğrulama tekniği (10-fold cross validation) kullanılarak hesaplanmıştır.

4 Bulgular ve Tartışma

Bu çalışmada zararlı ve zararsız URL'lerden elde edilen 79 özellikli bir verisetinde makine öğrenmesi yöntemlerinden Rastgele Orman, AdaBoost ve Gradyan Arttırılmış Karar Ağacı yöntemleri kullanılarak sınıflandırma işlemi gerçekleştirilmiştir.

Boyut azaltma işlemi temel bileşenler analizi yöntemi ile WEKA uygulaması kullanılarak uygulanmıştır ve sonucunda verisetini en iyi şekilde temsil edecek 10 ve 21 temel bileşen seçilmiştir.

Tablo 2'te 10-katlama çapraz doğrulama tekniği kullanılarak ikili ve çoklu sınıflandırma için elde edinilen doğruluk değerleri gösterilmiştir. 10 temel bileşen için %99.42 ikili sınıflandırma, %95.68 çoklu sınıflandırma ve 21 temel bileşen için %99.52 ikili sınıflandırma ve %96.56 çoklu sınıflandırma doğruluk değeri ile Rastgele Orman algoritması diğer iki algoritmaya göre daha yüksek performansa ulaşmıştır.

Tablo 2. Toplulukla öğrenme yöntemleri sonuçları

Algoritma	Temel Bileşen Sayısı	İkili Sınıflandırma		Çoklu Sınıflandırma	
		Doğruluk (%)	Model Oluşturma Süresi (sn)	Doğruluk (%)	Model Oluşturma Süresi (sn)
Rastgele Orman	**10**	**99.42**	**22.30**	**95.68**	**15.22**
Rastgele Orman	**21**	**99.52**	**32.55**	**96.56**	**20.00**
AdaBoost	10	81.03	1.53	36.62	0.11
AdaBoost	21	81.03	3.90	36.62	0.24
Gradient Boosting	10	86.42	5.20	79.93	20.98
Gradient Boosting	21	94.28	10.6	85.1	36.44

Tablo 3. Önerilen yöntem ve aynı veriseti üzerinde uygulanan farklı yöntemlerin karşılaştırılması

Model	İkili Sınıflandırma	Çoklu Sınıflandırma
Önerilen yöntem	%99.42	%95.68
[8]	%99	%93-99
[10]	%99	-
[12]	%90.18	-

21 temel bileşen ile ulaşılan model performansı daha yüksekken, 10 temel bileşen kullanılarak da daha kısa bir sürede yüksek bir sınıflandırma doğruluğuna ulaşılmıştır. Yükseltme yöntemlerinden AdaBoost ve Gradyan Arttırma algoritmaları ikili sınıflandırma için yüksek doğruluklar elde etmesine rağmen çoklu sınıflandırma problemi için kullanılabilecek uygun algoritmalar olmadığı sonucuna varılmıştır. 0.11 sn ve 0.24 sn ile en kısa model oluşturma süresine AdaBoost algoritması ulaşmasına rağmen, sınıflandırma doğrulukları daha yüksek olan Rastgele Orman ve Gradyan Arttırılmış algoritmaları 10 temel bileşen için 14 sn ve 20 sn gibi bir sürede model oluşturmuştur. Kullanılan üç sınıflama yöntemi içerisinde 21 temel bileşen ile ulaşılan model oluşturma süreleri ve doğruluk değerleri daha yüksektir.

Elde edilen sonuçlar aynı veriseti üzerinde farklı yöntemler uygulayan üç farklı [8, 10, 12] çalışma ile de karşılaştırılmıştır. 10 temel bileşen kullanılarak oluşturulan model diğer çalışmalar ile karşılaştırılmak üzere seçilmiştir. Tablo 3' te önerilen yöntem ve karşılaştırılan çalışmalarda ulaşılan ikili ve çoklu sınıflandırma için elde edilen doğruluk değerleri gösterilmektedir. Aynı verisetini kullanan [8]'de CfSubSet özellik seçme algoritması ve Rastgele Orman yöntemi birlikte kullanılarak ikili sınıflandırmada yaklaşık %99 ve çoklu sınıflandırmada her bir sınıf etiketi için %93-99 arası, ortalama %97 doğruluk oranları elde edilmiştir. Model performansları sadece doğruluk değerine göre yapıldığından dolayı model oluşturma süresi ile ilgili bir bilgi bulunmamaktadır. [10]'da seçilen 28 özellik ile birlikte Karar Ağacı algoritması kullanılarak sadece ikili sınıflandırma problemi için 0.29 saniye işleme süresine ve %99 doğruluk değerine sahip bir model oluşturulmuştur. [12]'de ise Naive Bayes algoritması ile seçilen üç özellik kullanılarak bir ikili sınıflandırma modeli oluşturulmuştur ve oluşturulan model %90.18 ikili sınıflandırma doğruluğuna sahiptir. Karşılaştırılan çalışmaların üçünde de 10-katlamalı çapraz doğrulama tekniği kullanılmamıştır fakat önerilen yöntemde bu teknik kullanılarak farklı test ve eğitim verisetlerinde de aynı performansa ulaşıp ulaşılamayacağı test edilmiştir.

Bu çalışmada önerilen yöntem Rastgele Orman algoritması ile birlikte PCA boyut azaltmayı birleştirerek %99.42 model doğruluğu ile [8]'deki çalışmadan %0.42 daha fazla ikili sınıflandırma doğruluk değerine ulaşmıştır. Önerilen yöntem %95.68 çoklu sınıflandırma doğruluğu ile yaklaşık %97 çoklu sınıflandırma başarısı olan [8]'den %1.32 daha düşük bir performans göstermiştir.

[10]'da oluşturulan Karar Ağacı modeli 28 özellik kullanarak %99 doğruluk değerine ve 0.29 saniye işleme süresi ulaşırken bu çalışma kapsamında önerilen yöntem seçilen 10 temel bileşen ile birlikte %0.42 daha fazla doğruluk değerine ulaşmıştır. Önerilen yöntemde [10]'daki modelden farklı olarak 10-kat çapraz doğrulama tekniği kullanılarak model oluşturma işlemi gerçekleştirildiğinden dolayı yaklaşık 22 saniye daha uzun bir model oluşturma süresine sahiptir. Önerilen yöntem Naive Bayes kullanılarak oluşturulan ve %90.18 doğruluk

değerine ulaşan [12]'den %9.24 daha yüksek bir ikili sınıflandırma sonucu elde edilmiştir. Diğer çalışmalar ile kıyaslandığında bilgi kaybına sebep olabilecek özellik çıkarma yöntemlerinin aksine en az bilgi kaybıyla boyut küçültmek için yapılan PCA yöntemi kullanılarak da %99.42 ikili sınıflandırma ve %95.68 çoklu sınıflandırma değerine ulaşılabileceği gösterilmiştir.

5 Sonuçlar

Çalışmada ISCX-URL 2016 veriseti üzerinde üç farklı toplulukla öğrenme yöntemi kullanılarak zararlı ve zararsız URL'lerin sınıflandırılma işlemi gerçekleştirilmiştir. Verisetinde toplam 79 nitelik bulunmaktadır. Verisetindeki çok boyutluluk problemine çözüm olarak PCA boyut azaltma yöntemi kullanılarak model oluşturma süresi ve depolamada avantaj sağlanmıştır. PCA boyut analizi ile yüksek korelasyona sahip olan niteliklerin verisetinden kaldırılması ile model oluşturmak için kullanılacak olan veriseti dosya boyutunda azaltma olmuştur. Örneğin verisetinin PCA analizinden önceki boyutu 14 MB iken 10 temel bileşen ile ikili sınıflandırma için kullanılacak olan veriseti 5.46 MB ile daha düşük bir değere ulaşmıştır.

10-katlamalı çapraz doğrulama tekniği kullanılarak ikili sınıflandırma probleminde ortalama %99.42 ve çoklu sınıflandırma probleminde ortalama %95.68 doğruluk değeri elde edilmiştir. Zararlı URL tespitinde farklı çalışmalarda da kullanılan ve iyi bir performans gösteren toplulukla öğrenme yöntemleri kullanılarak yüksek doğruluk değerlerine ulaşılabileceği gösterilmiştir. Ayrıca torbalama yöntemlerinden olan Rastgele Orman algoritması yükseltme yöntemlerine göre daha yüksek bir doğruluk değerine ulaşmıştır.

Kaynaklar

[1] Dwan Jr, Robert A., Alex M. Tavares. Predictive Analysis: Machine Learning Models for URL Classification. Diss. Worcester Polytechnic Institute, 2019.

[2] Gartner. "Gartner Forecasts Worldwide Security and Risk Management Spending to Exceed $150 Billion in 2021". https://www.gartner.com/en/newsroom/press-releases/2021-05-17-gartner-forecasts-worldwide-security-and-risk-managem (8.12.2021).

[3] McGrath DK, Gupta M. Behind Phishing: An Examination of Phisher Modi Operandi. LEET.15, 8-4, 2008 Apr.

[4] Ma J, Saul LK, Savage S, Voelker GM. "Identifying suspicious URLs: an application of large-scale online learning". *Proceedings of the 26th annual international conference on machine learning*, 681-8, 2009.

[5] Thomas K, Grier C, Ma J, Paxson V, Song D. "Design and Evaluation of a Real-Time URL Spam Filtering Service". *2011 IEEE Symposium on Security and Privacy*, 447-462, 2011.

[6] Choi H, Zhu BB, Lee H. "Detecting malicious web links and identifying their attack types[C]". *Usenix Conference on Web Application Development*, 11-11, 2011.

[7] Lin MS, Chiu CY, Lee YJ, Pao HK. "Malicious URL filtering—A big data application". *Proc. IEEE Int. Conf. Big Data*, 589-596, 2013.

[8] Chu W, Zhu BB, Xue F, Guan X, Cai Z. "Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing urls". *2013 IEEE International Conference on Communications (ICC)*, 1990-1994, 2013.

[9] Mamun MS, Rathore MA, Lashkari AH, Stakhanova N, Ghorbani AA. "Detecting malicious URLs using lexical analysis". *Proc. Int. Conf. Netw. Syst. Secur*, 467-482, 2016.

[10] Joshi A, Lloyd L, Westin P, Seethapathy S. "Using lexical features for malicious url detection – a machine learning approach". 2019.

[11] Powell A, Bates D, Van Wyk C, de Abreu D. "A crosscomparison of feature selection algorithms on multiple cyber security data-sets". Stellenbosch University, 2019.

[12] Singhal S, Chawla U, Shorey R. "Machine Learning & Concept Drift based Approach for Malicious Website Detection". *2020 12th International Conference on Communication Systems & Networks*, 582-585, 2020.

[13] Wang S, Wang Y, Tang M. "Auto Malicious Websites Classification Based on Naive Bayes Classifier". *In 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 443-447, 2020

[14] Sahingoz OK, Buber E, Demir O, Diri B. "Machine learning based phishing detection from urls". *Expert Syst. Appl*, 117, 345-357, 2019.

[15] Canadian Instıtute for CyberSecurity, "URL Dataset (ISCX-URL-2016)", https://www.unb.ca/cic/datasets/url-2016.html, 2016, (5.5.2020).

[16] Principal Component Analysis Tutorial, https://www.dezyre.com/data-science-in-

pythontutorial/principal-component-analysis-tutorial, (20.6.2020).

[17] Sun Y, Zhang H, Zhao T, Zou Z, Shen B, Yang L. "A new convolutional neural network with random forest method for hydrogen sensor fault diagnosis". IEEE Access, 8, 85421-85430, 2020.